California Department of Transportation (Caltrans)
Division of Transportation Systems Information
Office of Travel Forecasting & Analysis
1120 N Street
Sacramento, CA 95814

***

DMV DATA ANALYSIS:
Joint Project of
CEC, CALTRANS, & PARKS

***

DMV Vehicle Registration File Passes
Processed Under Interagency Agreement No. 72A0043:
10-04 & 04-05

***

July 12, 2005

Project Manager: Gareth Occhiuzzo,
CEC
916-654-4779

Contract Manager: Soheila Khoii,
Caltrans
916-651-7578

Technical Consultant: Robert Cenzer,
RC Consulting

MEMORANDUM OF UNDERSTANDING
Between the
Department Of Motor Vehicles & California Energy Commission

MOU No. 600-03-100 was executed on 10-12-04 between the DMV and the Commission. This document announces the DMV's endorsement of the Commission's DMV Data Processing Methodology along with other unprecedented privileges.

[Excerpt]

"This memorandum of understanding, hereinafter referred to as MOU, is between the Department of Motor Vehicles, State of California, hereinafter referred to as DMV, and the California Energy Commission, hereinafter referred to as Commission, for the purpose of establishing an ongoing agreement between the DMV and the Commission to perform Vehicle Registration (VR) ad hoc requests for government entities. The Commission's legislatively mandated transportation energy demand and supply responsibilities require a level of detailed vehicle population information not readily available from the DMV or private entities. A 1991 MOU between the DMV and the Commission resulted in the Commission's purchase of vehicle identification number (VIN) decoding software for the DMV, and the DMV's delivery of semi-annual "snapshots" of its VR database to the Commission. The Commission has since developed a VR processing methodology that transforms the partially decoded VR records of each DMV snapshot into a variety of refined vehicle population data to meet the Commission's objectives and those of its funding partners. Based upon ten years of staff-level interagency cooperation and positive feedback from requester referrals, *the DMV recognizes the Commission's VR data processing methodology as a valuable source of vehicle population information* when time or resource constraints, or a state or national emergency, may otherwise limit DMV's dissemination of data to government requesters. Under this voluntary MOU and by referral of the DMV, the Commission agrees to use its VR processing methodology to provide processed VR data to government entities. The Commission shall be entitled to receive full cost reimbursement from such requesting entities including, but not limited to, consultant fees, data processing fees, and administrative overhead costs. Finally, the Commission reserves the right to unilaterally decline requester referrals from the DMV."

TABLE OF CONTENTS

JOINT-AGENCY
DMV DATA PROJECT

History

1. **Need for Vehicle Population Data**: Vehicle population information is a key ingredient in a variety of transportation analyses. Inaccurate vehicle counts can significantly distort state research and planning for energy requirements, infrastructure needs, and air quality issues.

2. **Limitations of Available Vehicle Data Sources**:
   - **A. DMV**: DMV maintains a limited public reporting system, as its legislative mandate is focused primarily on the processing of driver license data in support of law enforcement agencies and the court system. Moreover:
     1) **Type-License vs. Size Class**: DMV's public data reports are based mostly on type-license registration categories (e.g., passenger, commercial, government, etc.), while the Joint Project has built and maintains an extensive guidefile that permits translation of vehicle make/series/model parameters into standard auto industry size classes (e.g., subcompact car, midsize cross-utility, standard pickup). The latter vehicle classification scheme is often more desirable for transportation analyses (see Table 12: Vehicle Count Comparisons, pp. 29-33);
     2) **Fleet Use**: DMV's public data reports do not categorize vehicle registration data by vehicle usage. The Joint Project's processing methodology can distinguish vehicles by primary usage categories (e.g., personal, business, daily rental, government), and by fleet size (i.e., fleets of 1 to *n* vehicles); and
     3) **Unidentifiable Vehicles**: DMV's public data reports do not resolve unknown vehicle identification issues [e.g., invalid and duplicate Vehicle Identification Numbers (VINs), unassigned Gross Vehicle Weight Ratings (GVWRs), erroneous fuel-type designations)]. The Joint Project's methodology resolves a majority of these errors.
   - **B. R.L. Polk**: Due to the per record charge by DMV ($.10 per record @ 46 million records), Polk purchases only the more recent model-year records. Consequently:
     1) **New Vehicle Sales**: Polk's focus is new vehicle sales, rather than the entire vehicle population. However, California's temperate climate underscores the importance of older vintage information;
     2) **NCRVs**: Polk does not purchase and therefore does not analyze records identified *as not-currently-registered vehicles* (NCRVs). However, record-level analysis by the Joint Project (discussed subsequently) revealed that NCRVs compose approximately 5-7 percent of the *operational* vehicle fleet (See Table 12, p. 29);
     3) **Standard vs. Custom Reports**: Due to their focus on new vehicle sales, Polk's standard reports offer limited coverage. Its custom reporting service is expensive. Per discussions with Polk staff, the product slate generated by the Joint Project would price out at several hundred thousand dollars per year; and
     4) **Limited Analysis of Unidentifiable Vehicles**: Polk's standard products offer limited analysis of problem records [e.g., invalid VINs can preclude vehicle identification of any number of important variables, including vehicle type (e.g., car versus truck), fuel-type, and truck size].

3. **Formation of Joint-Agency Project**: Due to the importance of accurate vehicle population data in transportation analysis, and the lack of a viable alternative, in 1991, CARB and the CEC co-funded the purchase of R.L. Polk's VINA software for DMV. This software permits DMV to translate a vehicle record's VIN into valuable information (make, series, model, model year, fuel type, GVWR, engine size, etc.). DMV keeps this software current with continuous updates from Polk. In exchange, the DMV provides annual copies of the VR file to the Joint Project. Caltrans became a funding partner a couple of years later, and the arduous task of developing a comprehensive processing methodology has been ongoing since that time.

4. **Users of Joint-Agency's DMV Data Products**: During FY-04/05, the Joint Project logged approximately 85 data requests. Most are from California government agencies or their consultants and stakeholders. Among these were more than a dozen from the DMV itself. Current and past requesters include: Los Angeles County Metropolitan Transportation Authority, Dept. of General Services, UC-Berkeley, San Francisco State University, Dept. of Insurance, Bay Area Air Quality Management District, Bay Area Metropolitan Transportation Commission (MTC), San Diego Association of Governments (SANDAG), Southern California Association of Governments (SCAG), Board of Supervisors for the Counties of San Francisco and Marin, Ford Motor Company, National Highway Transportation Safety Administration (NHTSA), U.S. Census Bureau, and the CA Attorney General.

Ongoing Nature

1. **Ongoing Need for Programming/Analysis**: The following information describes why a push-button processing program cannot be created that would obviate the need for ongoing programming/analysis of each successive DMV VR data base.
   - **A. Size of Data Base**: With the possible exception of the Franchise Tax Board data base, DMV's VR data base is the largest in California's state government system. Each DMV VR data base contains 25 gigabytes of vehicle information representing approximately 46 million records. The length of each record is 527 characters and includes 55 data fields housing 85 variables. The Commission receives two such data bases from DMV annually (October and April).
   - **B. Quality of Raw Data**: Unfortunately, the DMV data base contains nearly as many data problems as raw data. At present, the Data Project's processing methodology consists of 55 steps whose core function is to algorithmically identify missing, and correct existing information. While some steps are manually performed, many are coded in SAS. But, virtually all steps require a manual support process to address hundreds of thousands (down from millions) of record "exceptions"; these are records with unique problems that cannot be algorithmically resolved (i.e., they require custom attention). (To illustrate, Section 1.C discusses one of many data issues that must be addressed manually.) Over time, we have and continue to automate and streamline the process. However, the scope and nature of changes from data base to data base essentially preclude elimination of a manual support process for the foreseeable future.
   - **C. Vehicle Identification Issues:**
     1) **Pre-1981 Vehicles**: A VIN is the primary source of vehicle information (e.g., make, series, model, model-year). In the late 1970s, NHTSA standardized the VIN format to 17 characters for model years 1981 and newer. This change was adopted by nearly all vehicle manufacturers. However, California's temperate climate fosters a sizeable number of pre-1981 vehicles – around 3 million in the October 2002 VR data base. Though this count is

slowly declining, manual identification of this large vehicle segment at the Make-Series-Model level is a long, labor intensive process.

2) **Invalid Vehicle Identification**: Partially corrupted VINs inhibit automated identification of a vehicle record. The work-around is a labor intensive manual process using third-party VIN decoding software.
3) **Fuel Type Mis-identification**: This ongoing problem results from the relatively uncoordinated efforts of several parties responsible for encoding a vehicle's fuel type in the VIN. Having worked with all three parties (the auto manufacturers, R.L. Polk--a primary VIN decoding software company, and DMV), it is our opinion that this problem is likely to continue indefinitely.
4) **Double Counted Records**: Each VR data base contains a unique set of double counted vehicles that must be identified, analyzed, and appropriately resolved.
   a. RIPs: At any point in time, 2-3 million vehicles fall into DMV's Registration in Process (RIP) status. This results from critical omissions by the applicant during re-registration (e.g., missing proof of insurance or smog certification).
   b. VINs: Each data base contains thousands of duplicate VINs that require manual effort to resolve.

D. Evolution of Raw Data:

1) **Record Changes**: DMV processes 100,000–150,000 vehicle transactions per day that impact millions of data base records annually. Most transactions are key-entered resulting in changes (and often errors) to millions of records annually. Transaction examples include: Migration from out-of-state, change in registration status from operational to non-operational or junk, name and/or address change; vehicle resale, new vehicle purchase. Note that with a 46 million record data base, a small data entry error of only 1% results in 460,000 errors.
2) **Record Additions**: Approximately 2 million new vehicles of all sizes are purchased in California annually. Consequently, DMV manually enters 2 million new vehicle records into its VR data base each year.
3) **Auto Industry Updates:**
   a. The number of unique vehicle Make, Series, and Model combinations in the 10-02 data base was 4,614.
   b. The number of unique vehicle Make, Series, Model, and Model-Year combinations in the 10-02 VR data base was 23,053 (up 23% from 18,668 in 10-00).
   c. Each October data base contains hundreds of changes/additions to the Make, Series, and Model combinations that must be extracted and manually mapped to a corresponding EPA-type size class. Once this manual procedure is accomplished, the processing software then can use this template to automatically compile vehicle counts according to size class groups needed for analyzing a particular transportation issue or for input to transportation forecasting models.

E. Continuous Need for New Information by Requesters Requires Ongoing Project Programming and Analysis. Some recent requests from the project's log follow:

1) Vehicle counts by new weight subcategories [e.g., segmenting the industry weigh standards from (0-6000 and 6001-10,000) to (0-5000, 5001-8500, and 8501-10,000)].
2) Vehicle counts compiled by changing CARB emissions categories (e.g., PZEV, SULEV, LEV II, MDV-SULEV).
3) New size class mappings for new make-series-model industry offerings (e.g., breakout of new cross-utility series from traditional SUV class series).
4) New vehicle types: Hybrids, NEVs, and Hydrogen/New Diesel when available.

5) Identification of light-duty vehicles at "model" level (i.e., breakout beyond make and series).
6) Identification of medium-heavy-duty vehicles at "series" level (i.e., breakout beyond make).
7) Off-highway vehicles: motorcycles, snowmobiles, ATVs, etc.
8) County to zip code mappings: resolve discrepancies in DMV's address fields.
9) Special extractions: e.g., UPS's CNG fleet.

2. **Current/Future Objectives of the Project:**
   - **A. Mine New Areas of the Data Base**: For example, identification of gaseous-fueled vehicles—Propane, LPG, LNG; identification of buses; make-series identification of heavy-duty vehicles; weight identification of medium/heavy-duty vehicles; identification of vehicles by emission categories.
   - **B. Increase Accuracy of the Current Processing Algorithms**: For example, distinguishing 1, 2, and 3 vehicle fleets by personal vs. commercial use.
   - **C. Code the Manually Computed Support Processing Steps**: Currently done manually due to non-standard nature of analysis.
   - **D. Automate the Processing Methodology to the Greatest Extent Possible**: For example, automatic update of guidefile matrix that translates make-series-model-modelyear into auto industry size classes.
   - **E. Expand the Documentation of the Processing Methodology**: Expand existing documentation to describe the operation and maintenance of the software routines.
   - **F. Expanded Report Writing Routines**: For example, expand transportation regions to individual counties.
   - **G. Port the Processing System from Teale to In-House Servers**: This would increase processing efficiency and provide cost savings for the project.

Data Processing Methodology

1. **Two DMV File Passes Per Year**: Two DMV VR files are processed each year: October 1, and April 1. The October file pass provides the raw information for the ***currently registered*** vehicle analysis. The April file pass (of the next calendar year) is used to identify which of the ***not-currently registered vehicles*** (NCRVs) in the October file pass have *re-registered* as of April. The merged results of both file passes yield a production data base of ***operational*** vehicle counts.

2. **Processing Methodology:**
   - **A. Identifying Vehicles – Key Issues:**
     - **1) Size of Raw Data Base**: Each VR data base contains 46-million records consisting of 527 characters across 75 variables, or about 25 gigabytes of information per file pass (see Table 1: Record Layout, pages 1-5). Unfortunately, there are nearly as many data problems as there are raw data. For obvious reasons, the raw data are cleaned and compiled algorithmically, to the greatest extent possible. Consequently, the Data Processing Methodology developed by the Joint Project is quite complex. The 55 main SAS programs that compose the software process are listed and flow charted in Table 2: Program and Data Flow Chart, pages 6-11.

The core function of the data processing methodology is to identify missing information, and to create new identification variables. All the intermediate results and final results are contained in new data files generated by the SAS codes. The Flow Charts (in Parts A-D, pages 8-11) present the sequence of the SAS programs, and the relationships between programs and data groups. The top row of the chart identifies the SAS program. Under each program heading, the text boxes in the column present the data files generated by that program. These data files may be used as inputs in later program steps. Data flow is indicated by arrowed lines. A brief program description is provided near the bottom of each program column. Key variables generated in each program step also are listed at the bottom of the chart.

As the flow charts indicate, some data steps are performed or supported by manual processes. Typically, the focus of such steps is record exceptions, or problematic records that cannot be algorithmically processed. Such records are identified and segregated into homogeneous issue groups. Then, the arduous and time consuming task of crafting custom solutions is performed, thus resolving the identified issues to the greatest extent possible. The Joint Project has and continues to compile information from a variety of sources for this purpose. Sources include the CEC, CARB (e.g., list of EV and NEV manufacturers and their corresponding vehicle issues from Krista Eley), third-party VIN-decoding software, vehicle technology reference manuals, and the like. Total automation of the complete processing methodology is an ongoing goal of this project, as funding becomes available.

2) **DMV Vehicle Categories**: DMV segments the raw VR data into several non-descript categories that are not especially useful for various kinds of transportation analyses. They include: Passenger, Truck, Motorcycle, Vessels, Off-Highway, and Unknown. A detailed size class system more meaningful for transportation analysis was developed by the Joint Project. Dubbed the Guidefile Data Base, it is described below in more detail.
3) **Operational/Nonoperational Vehicles**: The non-descript vehicle categories are cross-stratified by several DMV registration conditions that must be identified and interpreted in order to distinguish operational from non-operational vehicles. The cross-stratification categories include: Junk (e.g., non-revivable, salvage), Registration-in-Process (RIPs), Planned Non-Operational (PNOs), Legally Protected (e.g., Judges), Lien Sales, Not In Service (e.g., surrender of title), etc.
4) **Unidentifiable Vehicles**: VINs were standardized by the National Highway Transportation Safety Administration (NHTSA) for model-years 1981 and later. Polk's VINA software is able to transform the VINs of many of these records into meaningful information. However, such is not the case for many pre-1981 records and for invalid VINs. In some cases, the missing information can be drawn from DMV's manually entered data fields (e.g., Body-Type, Type-Body, Type-Vehicle, Body-Style), which have been added to the file pass at the request of the Joint Project. However, the balance must be interpreted by manual procedures or algorithms developed by the Joint Project.
5) **Duplicate Vehicles**: Duplicate VINs, a non-trivial problem, must be identified and resolved.
6) **Currently/Noncurrently Registered Vehicles**: Most file passes contain approximately 5 million records identified as *not currently registered vehicles* (NCRVs). Among reasons for lack of current registration:
   a. the vehicle is no longer in use, but DMV has not yet purged the record from the data base;
   b. the vehicle is in use, but not registered due to:

1) routine difficulties in re-registering: e.g., insurance, smog check;
2) special owner problems--financial/other;
3) DMV processing delays.

NCRV records from the October file pass must be identified, segregated, and matched against the April filepass to determine which have re-registered. These then become part of the *operational* count of vehicles.

Note: Use of the April file pass for determining operational counts is based on two studies of NCRV records by RC Consulting. These studies tracked the change in registration status of not currently registered vehicles 24-months forward from a base point. Both studies revealed that the preponderance of NCRVs that ultimately reregistered do so within 6 months of the base file pass. As opposed to sample-based studies that have reported NCRVs at 1-3 percent, these record-specific vehicle population analyses revealed that NCRVs compose approximately 5-7 percent of the *operational* vehicle fleet.

B. Assigning Light-Duty Vehicle Classes: Completion of the various VIN-decoding and related vehicle identification steps described above provides a specific Make (e.g., Ford), Series (e.g., Taurus), Model (e.g., LX), and Model-Year designation for each light-duty vehicle record in the VR data base. The next step is translating these parameters into industry-standard size classes (e.g., subcompact car, midsize cross-utility, standard pickup) by model-year. Currently, more than 23,000 Make/Series/Model/Model-Year combinations exist. The Joint Project developed an algorithmic approach that accomplishes this translation. The Guidefile (see Table 3—Guidefile Data Base, page 12), a large two-dimensional matrix, contains a column for each model-year, and a row for each of the make/series/model combinations. The cells of this large matrix identify the 15 size class designations currently in use by the Joint Project (see Table 4—Size Classes, and Table 5—Size Class Definitions, pages 13-15). The Guidefile is updated with each successive file pass to reflect any new makes/series/models that are produced by the vehicle manufacturers, or any size class changes to existing makes/series/models.

C. Assigning Medium/Heavy-Duty Vehicle Classes: Composed mostly of trucks, these records are further distinguished by (1) Gross Vehicle Weight Ratings (GVWRs) (see Table 6—GVWR, page 16), and (2) approximately 40 body-styles (see Table 7—Heavy-Duty Body Styles, page 17). The non-trivial incidence of conflicts between vehicle weight and GVWR values contained in the VR file pass are resolved by subroutines and support analysis (e.g., third-party VIN-decoding software, vehicle technology reference manuals) developed and compiled by the Joint Project.

D. Identifying Fuel Types: The VR file pass provides raw fuel-type information from two sources—(1) VINA, based on VIN decoding, and (2) DMV, based on registration input. Often, these two sources conflict, or are completely erroneous. Such problems are resolved to the greatest extent possible by encoded search-and-translate subroutines based on known Make/Series/Model/Model-Year information collected from a variety of sources including the CEC and CARB, third-party VIN-decoding software, and other support analysis developed and compiled by the Joint Project. Vehicle counts for eight fuel-type categories are currently available: Gasoline, Diesel, Electric, Neighborhood Electric, Hybrid, Flex-Fuel Alcohol, Total Alcohol, Compressed Natural Gas, and Other Gaseous.

E. Identifying Vehicle Fleets:

1) Names/Addresses: In support of fleet identification, owner names and addresses (contained in 4 fields totaling 120 characters) are standardized, and county versus zip code conflicts are resolved.
2) Fleet Size: Vehicle fleets of sizes 1 through n are identified on the basis of common ownership and location. Search keys composed of combined segments of the standardized

name and address fields are created and applied in the initial pass. A second identification pass based on an exact address match provides the final fleet identification results.

3) Fleet Use: Identified fleets are further distinguished by four fleet usage categories: (1) Government (by type-license codes and owner name fields); (2) Daily Rental (by a master list of all daily rental companies); (3) Other Commercial, and (4) Personal. The latter two categories are distinguished on the basis of a complex algorithm that extracts and applies business-identifying words from the owner name fields (e.g., Inc., Corp., Company, Associates, & Sons, Painting, Plumbing, Landscape, Bakery). (See Table 8—Fleet Use Categories, page 18)

F. Identifying 4-Wheel Drive (4-WD) Vehicles: 4-WD identification for model-years 1981 and later is contained in the VIN. 4-WD identification for most pre-1981 vehicles must be derived based on Make/Series/Model/Model-Year data for each manufacturer. The Joint Project compiled and encoded into subroutines the information necessary to identify all 4-WD vehicles contained in the VR file pass. These search routines are conservatively constructed to yield counts with a very high level of confidence.

G. Identifying New Vehicle Sales: Following identification of vehicle records at the Make/Series/Model/Model-Year level, the processing methodology then identifies which vehicle records were first sold new during the 6-month window of a given VR file pass. This is accomplished using the Last Ownership Issue Date, which determines each vehicle's specific date of sale.

Note: Though the VR File Passes are static snapshots, they reflect virtually a complete picture of CA's vehicle population during the past 6 months. Vehicle records would be missing only if they meet BOTH of the following criteria: (1) a vehicle left the state or was destroyed by accident, AND (2) DMV purged the registration record of such vehicles during the past 6 months. Satisfaction of the second criteria is highly unusual -- expired vehicle records typically linger in the VR data base for 48 months or longer. Faster record purges typically occur only if: (1) the vehicle was part of a commercial fleet (e.g., daily rental) that was traded out of state, or (2) title of a vehicle destroyed by accident was transferred to a licensed junkyard in CA. In the latter case, the vehicle record would be included in the junked records (reporting is required by law) that are a part of each file pass.

3. Reporting Methodology:

A. Custom Reports: The completion of all processing steps results in a SAS "production" version of the "raw" filepass delivered by DMV. The production version adds about 150 characters of data to each record. Once transformation of the raw file pass to a processed data base has been achieved, the marginal cost of preparing custom reports is quite low for data fields included in the transformation process. Custom tabular reports are easily generated by SAS routines created for that purpose. The Joint Project can report vehicle counts down to the zip code level. Under a recent specially funded GIS project, personal and commercial fleet counts were reported down to the Census Block level for several counties.

B. Standard Production Reports: The Joint-Agency DMV Data Project has designed a Report Generator Methodology that packages key results into two easy-to-use Excel programs (see Tables 9 & 10 – Light-Duty & Medium/Heavy-Duty Vehicle Report Generators, pages 19-20) that can be run on a standard Pentium PC or Laptop. The Light-Duty Program alone is capable of generating 16,200 vehicle reports that can be customized by the user, based on a broad range of parameters that include:

- 4 categories of geographic regions (e.g., 12 Caltrans Planning Districts, 6 Metropolitan Planning Organization Regions; 58 Individual Counties plus Out-of-State; and 5 CEC Transportation Regions). See Table 11 – County Composition of Geographic Regions, page 21-28.
- 12 fuel-type categories;
- 18 fleet-type categories (e.g., personal, daily rental, other commercial, city government, state government);
- 9 fleet-size categories (e.g., fleets of 1-9, 10-19);
- 15 light-duty vehicle size classes [e.g., subcompact car, standard pickup, heavy vans (8,501-10,000)];
- 6 heavy-duty vehicle gross-weight categories (i.e.,, Gross Vehicle Weight Ratings 3-8);
- 17 individual model years (e.g., 2005-1989) and 1 lump-year (1988 and earlier); and
- 41 medium/heavy body styles (e.g., ambulance, bus, garbage, motorized home).

SUM2004.xls addresses light, medium, and heavy-duty vehicle counts by fleet-size, while MHSUM2004.xls reports body-style and gross weight information for medium and heavy-duty vehicles. Both programs are structured as a multi-tiered data sandwich: The first level is the calculator sheet, the middle are the data sheets (1 to 8 hidden worksheets), and the bottom defines the user-input codes. These levels are selected by clicking on the tabs located at the bottom of the screen. The user simply inserts desired input codes into the green highlighted boxes at the top of the calculator sheet, and pushes the [F9] key to calculate the requested results. All output pages are conveniently formatted as 8.5 x 11.

C. New Production Reports:

1) In addition to the Excel Report Generator program described in 3.B that provides operational counts based on the combined October and April DMV VR data bases, a Light-Duty Report Generator is now available that provides similar output for the October data base alone. As such, these results of necessity reflect Currently Registered and not Operational vehicle counts. (Recall that operational counts from the October data base typically are 4-7% higher when they include re-registrations from the April data base, which reflects a more accurate picture of what is actually on the road.)
2) A second new report, the County-Zip Code Excel Report, provides currently registered vehicle counts for each zip code within each of CA's 58 counties. These vehicle counts are reported by the aforementioned size classes, fuel types, and individual model-years. As previously discussed, usage info at the zip code level currently is unavailable as it would require record-level (as opposed to algorithmic) analysis which is beyond the scope of the current budget.

D. Future Direction:

1) Moving From Algorithmic To Record-Level Fleet Identification: This methodology shift likely holds the key to increased accuracy and comprehension for identification of vehicle fleets, especially heavy-duty fleet types, at the county level and ultimately for fleet-id expansion to the zip code level. Relaxation of limitations inherent in the existing algorithmic approach will be labor-intensive. Examples include permitting vehicle members of legitimate fleets to cross county boundaries as appropriate, collapsing inadvertently-split fleet sites, retrieving inadvertently "orphaned" fleet records (to correct fleetsize undercounting), and tracking the correct locations of inadvertently "adopted" fleet records (to correct fleetsize over counting). Additional harvesting of information contained in each Vehicle Record to identify fleet usage also is possible. Finally, based on past fleet identification analysis of this type, results for larger fleet sizes typically prove more reliable

than those for smaller fleet sizes due to a myriad of factors (e.g., higher commercial content, lower address complexity--no apt. complexes).

2) DMV's Axle Data Clean-Up: While potentially useful to Caltrans, the current state of this DMV data field offers little value. Third-party research will be necessary to identify axle configurations for all desired heavy-duty vehicle GVWR classes.